Student-teacher training with diverse decision tree ensembles

Jeremy H. M. Wong and Mark J. F. Gales

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, England
jhmw2@cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Student-teacher training allows a large teacher model or ensemble of teachers to be compressed into a single student model, for the purpose of efficient decoding. However, current approaches in automatic speech recognition assume that the state clusters, often defined by Phonetic Decision Trees (PDT), are the same across all models. This limits the diversity that can be captured within the ensemble, and also the flexibility when selecting the complexity of the student model output. This paper examines an extension to student-teacher training that allows for the possibility of having different PDTs between teachers, and also for the student to have a different PDT from the teacher. The proposal is to train the student to emulate the logical context dependent state posteriors of the teacher, instead of the frame posteriors. This leads to a method of mapping frame posteriors from one PDT to another. This approach is evaluated on three speech recognition tasks: the Tok Pisin and Javanese low resource conversational telephone speech tasks from the IARPA Babel programme, and the HUB4 English broadcast news task.
Index Terms: Student-teacher, random forest, ensemble, speech recognition

1. Introduction

In Automatic Speech Recognition (ASR), hardware limitations can often make it impractical to implement large models, even though they may perform well. Student-Teacher (ST) training [1] is a possible solution to this issue, by training a single student model to emulate the behaviour of the larger teacher model or ensemble of teachers. Only this single student needs to be used at test time. In ST training, there are many ways of propagating information from the teacher(s) to the student [2, 3, 4]. In a hybrid Neural Network-Hidden Markov Model (NN-HMM) acoustic model architecture, a common method of propagating frame posterior information is to minimise the KL-divergence between frame posteriors, represented by the NN outputs [2]. However to date, ST methods have assumed that the NN outputs of all models have identical interpretations, otherwise the KL-divergence criterion does not make sense. In ASR, this excludes the possibility of using different sets of state clusters between models.

One situation where different sets of state clusters are required is in a Random Forest (RF) ensemble [5]. Ensemble methods [6, 7] in general make a Monte Carlo approximation to Bayesian inference, by performing combination over a finite number of models. The ensemble captures the uncertainty about the model parameters that is encapsulated within the diversity of the models used. Methods such as Dropout [8], Adaboost [9], and using multiple Random Initialisations (RI) [4] produce a diversity of models within a fixed architecture. A diversity of architectures can be achieved by combining different model types [10]. Methods such as RF provide an additional mode of diversity by using a variety of Phonetic Decision Trees (PDT) [11], thereby capturing uncertainty about the state clustering. These different ensemble methods can be used concurrently to obtain a richer ensemble.

Ensembles often outperform single models [6], but can be computationally expensive to use. During decoding, it is more computationally efficient to combine the ensemble at the frame level [10, 12] than the hypothesis level [6, 7], as this only requires the processing of a single lattice for the whole ensemble. To further reduce the computational demand when using an RF ensemble, a multi-task architecture can be used [13], where all hidden layers are shared between models and only separate outputs are used for each PDT. The data then only needs to be fedforward through the hidden layers once. However, decoding a single student is less computationally demanding.

This paper extends the capability of ST training to allow different sets of state clusters. The proposal is for the student to emulate the logical context posteriors of the teacher, rather than the frame posteriors. This leads to a method of mapping the frame posteriors from one PDT to another. As such, ST training can be used to compress an RF ensemble. This also introduces the freedom to choose the output complexity of the student independently of the teacher.

2. Ensemble with different state cluster sets

It has been found that the acoustic representations of phones are strongly affected by their neighbouring contexts [14], leading to the use of context dependent phones. However, independently modelling all contexts requires too many trainable parameters to robustly estimate. To reduce the number of trainable parameters, similar contexts can be clustered together, with their observation likelihoods tied [14]. This can be achieved using a PDT [14], $\mathcal{T}$, which defines a many-to-one mapping from logical context HMM states, c, to physical state clusters, s, at the PDT leaves,

$$s_c = \mathcal{T}(c). \tag{1}$$

Finding a globally optimal PDT is a computationally intractable problem. As such, the PDT is usually trained by finding the greedy split at each node from a set of phonetically motivated questions. However, the resulting PDT is not guaranteed to even be at a local optimum of the total cost over the whole tree.

This training procedure can be modified to produce a variety of PDTs. In the RF method, diversity is achieved by injecting randomness into the node split selection. One way to inject randomness is to randomly select one of the top n splits at each node, instead of the greedy split [5]. Each PDT is associated with a separate NN [15], which learns to discriminate between its own state clusters. An ensemble of these models can approximately capture the uncertainty about the state clustering. Another way to obtain multiple PDTs is to explicitly train them to be different [16, 17]. These PDT forest methods can be used

concurrently with other ensemble methods, such as using different model types, to obtain a more diverse ensemble that better captures model uncertainty.

During decoding, the predictions of the models in the ensemble need to be combined together. Frame combination is more efficient that hypothesis combination, as only a single lattice needs to be processed for the whole ensemble. One approach to do frame combination of an RF ensemble is to convert the frame posteriors into observation pseudo likelihoods and take a linear average [12],

$$\widetilde{p}\left(\boldsymbol{o}_t\middle|c,\widehat{\boldsymbol{\Phi}}\right)=\sum_{m=1}^{M}\lambda_m\frac{P\left(s_c^m|\boldsymbol{o}_t,\boldsymbol{\Phi}^m\right)}{P\left(s_c^m\right)},\tag{2}$$

where $\boldsymbol{o}_t$ is the observation, t and m are the time and model indexes, M is the ensemble size, $\boldsymbol{\Phi}^m$ and $\widehat{\boldsymbol{\Phi}}$ designate the individual models and ensemble, and λ_m are the model interpolation weights that satisfy $\sum_m\lambda_m=1$ and $\lambda_m\geq 0$. Every logical context that gets mapped to the same set of M clusters, $\left\{s^1,\cdots,s^M\right\}$, will share the same combined pseudo likelihood. It is therefore possible to cluster these contexts together and tie their likelihoods without any loss of generality [12]. These clusters are referred to as RF tied states.

3. Student-teacher training

It can be computationally demanding to use a large model or an ensemble during decoding. One possible solution to this is ST training, where a single student model is trained to emulate the behaviour of the large teacher model or ensemble of teachers. During test time, only this single student model needs to be decoded, thereby reducing the computational demand. To learn from an ensemble where all teachers share the same state clusters, standard ST training propagates frame posterior information from the teachers to the student, by minimising the KL-divergence between their frame posteriors [2],

$$\mathcal{G}=-\sum_{t=1}^{T}\sum_{m=1}^{M}\alpha_m\sum_{s\in\mathcal{T}}P\left(s|\boldsymbol{o}_t,\boldsymbol{\Phi}^m\right)\log P\left(s|\boldsymbol{o}_t,\boldsymbol{\Theta}\right),\tag{3}$$

where T is the total number of training frames, α_m are the model interpolation weights such that $\sum_m\alpha_m=1$ and $\alpha_m\geq 0$, and $\boldsymbol{\Theta}$ designates the student model. It is also possible to interpolate forced alignments into the criterion target. However, it is shown in [4] that when the teachers have been sequence trained, the forced alignments do not benefit the student, and as such, shall not be used in this paper.

During decoding, the observation pseudo likelihoods only need to be computed from the student model, using

$$\widetilde{p}\left(\boldsymbol{o}_t|c,\boldsymbol{\Theta}\right)=\frac{P\left(s_c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right)}{P\left(s_c\right)},\tag{4}$$

instead of from each model in the ensemble.

4. Mapping posteriors between clusters

The teachers and student used with standard ST training are restricted to have the same state clusters, otherwise the KL-divergence of (3) cannot be used. This forbids the use of an RF ensemble. To allow for different sets of state clusters, a distance measure can be defined over the logical contexts. This paper proposes to train the student by minimising the KL-divergence between logical context posteriors,

$$\mathcal{F}=-\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}P\left(c|\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\log P\left(c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right),\tag{5}$$

where $\mathcal{C}$ is the set of all logical contexts. For simplicity, only a single teacher with a different PDT from the student is considered here. The student's logical context posteriors can be decomposed as

$$P\left(c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right)=P\left(c|s_c,\boldsymbol{o}_t,\boldsymbol{\Theta}\right)P\left(s_c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right).\tag{6}$$

Since the PDT is a deterministic mapping from logical contexts, c, to clusters, s, $P\left(c|s,\boldsymbol{o}_t,\boldsymbol{\Theta}\right)=0$ for all c that are not in cluster s. Substituting (6) into (5) leads to

$$\begin{aligned}\mathcal{F}=-\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}P\left(c|\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\Big[&\log P\left(s_c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right)\\&+\log P\left(c|s_c,\boldsymbol{o}_t,\boldsymbol{\Theta}\right)\Big].\end{aligned}\tag{7}$$

The student's NN weights in $P\left(s_c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right)$ need to be trained. The standard system does not have $P\left(c|s_c,\boldsymbol{o}_t,\boldsymbol{\Theta}\right)$, and it shall therefore be ignored in training. This simplifies the criterion to

$$\widetilde{\mathcal{F}}=-\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}P\left(c|\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\log P\left(s_c|\boldsymbol{o}_t,\boldsymbol{\Theta}\right).\tag{8}$$

It is inefficient to compute the sum over c, as there are many logical contexts. It is better to sum over state clusters in the student's PDT, $\mathcal{T}$, by re-expressing the criterion in the form of

$$\widetilde{\mathcal{F}}=-\sum_{t=1}^{T}\sum_{s\in\mathcal{T}}Q\left(s|\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\log P\left(s|\boldsymbol{o}_t,\boldsymbol{\Theta}\right).\tag{9}$$

This form can be obtained by expressing the target posteriors as

$$Q\left(s|\boldsymbol{o}_t,\boldsymbol{\Phi}\right)=\sum_{s^\Phi\in\mathcal{T}^\Phi}P\left(s\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)P\left(s^\Phi\middle|\boldsymbol{o}_t,\boldsymbol{\Phi}\right),\tag{10}$$

where

$$P\left(s\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)=\sum_{c:\mathcal{T}(c)=s}P\left(c\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right),\tag{11}$$

and s^Φ and $\mathcal{T}^\Phi$ are the teacher's state clusters and PDT respectively. When the student and teacher have the same PDT, $\mathcal{T}=\mathcal{T}^\Phi$, then the transformation reduces to the identity matrix, $P\left(s\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)=\delta\left(s,s^\Phi\right)$, leading to the standard ST criterion of (3).

It is the matrix transformation, $P\left(s\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)$, that makes it possible to do ST training across different PDTs, by mapping frame posteriors between these PDTs. However, standard ASR systems again do not yield $P\left(c\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)$. To address this, an approximation can be made that it is independent of the observation,

$$P\left(c\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\approx P\left(c\middle|s^\Phi\right).\tag{12}$$

The transformation will then also be independent of the observation, $P\left(s\middle|s^\Phi,\boldsymbol{o}_t,\boldsymbol{\Phi}\right)\approx P\left(s\middle|s^\Phi\right)$. This approximation loses some phonetic resolution. Computing the transformation then requires the estimation of $P\left(c\middle|s^\Phi\right)$, which can be expressed as

$$P\left(c\middle|s^\Phi\right)=\frac{P\left(c\right)}{\sum\limits_{c':\mathcal{T}^\Phi(c')=s^\Phi}P\left(c'\right)}\delta\left(\mathcal{T}^\Phi\left(c\right),s^\Phi\right).\tag{13}$$

It is possible to obtain $P(c)$ as a maximum likelihood estimate from forced alignments. To improve robustness, a discount factor can be incorporated into the estimate,

$$P(c) = \frac{N_c + \nu}{\sum_{c' \in \mathcal{C}} (N_{c'} + \nu)}, \quad (14)$$

where N_c is the number of times c appears in the forced alignments, and ν is the discount factor. This smoothing technique is commonly used in areas such as language modelling [18]. This allows $P\left(s\middle|s^{\Phi}\right)$ to be computed, which is used to map the teacher's frame posteriors to the student's state clusters. These mapped target posteriors can then be used with standard CE training infrastructures.

For an ensemble of teachers, the target posteriors can be computed as an average of the contributions from each teacher,

$$Q\left(s\middle|\boldsymbol{o}_t, \widehat{\boldsymbol{\Phi}}\right) = \sum_{m=1}^{M} \alpha_m \sum_{s^m \in \mathcal{T}^m} P(s|s^m)\, P(s^m|\boldsymbol{o}_t, \boldsymbol{\Phi}^m). \quad (15)$$

This proposed criterion thus allows ST training to be used together with an RF teacher ensemble, and for the student's PDT to be chosen independently of teachers' PDTs.

5. Experiments

The experiments were performed on the Kaldi speech recognition toolkit [19], and used the Tok Pisin (*IARPA-babel207b-v1.0e*) and Javanese (*IARPA-babel402b-v1.0b*) datasets, which are low resource tasks from the Babel programme [20], and the HUB4 English broadcast news (*LDC97S44* and *LDC98S71*) dataset. The Very Limited Language Pack (VLLP) was used for Tok Pisin, while for Javanese the Full Language Pack (FLP) was used, comprising approximately 3 hours and 40 hours of conversational telephone speech respectively. Graphemic lexicons [21] were used, along with trigram language models that were trained on the VLLP transcriptions for Tok Pisin and FLP transcriptions for Javanese. The standard 10 hours development sets were used for decoding. For HUB4, the 144 hours training set was used, comprising data from both the 1996 and 1997 releases. The standard phonetic lexicon was used, and the trigram language model was imported from the RT-04 system [22]. The 2.6 hours *Eval03* test set was used for decoding. Experimenting on these datasets allows an investigation of ST training over different performance ranges and lexicon types.

For all datasets, frame alignments were obtained from tandem Gaussian Mixture Model (GMM)-HMMs. These GMMs were trained on 107-dimensional multilingual tandem features [23] for Tok Pisin and Javanese, and 65-dimensional unilingual tandem features for HUB4. PDTs were trained with greedy splits, having 1000 leaves for Tok Pisin and 6000 leaves for both Javanese and HUB4. These greedy PDTs were used to construct RI ensembles. RF PDTs with the same number of leaves were trained, by randomly selecting a split from the best 5 at each node. Only splits that increased the likelihood beyond a threshold were considered. The alignments were mapped from the tandem GMM PDT to each of the PDTs. These alignments were used to train DNNs consisting of 4 layers of 1000 nodes for Tok Pisin, and 6 layers of 2000 nodes for both Javanese and HUB4. For Tok Pisin and Javanese, the DNN inputs consisted of the tandem features with a 9 frame context. For HUB4, the DNN input consisted of 40-dimensional filterbank features with first temporal derivatives and a 9 frame context. The DNNs were first initialised with layerwise pretraining using either the CE or ST criterion, and then fine-tuned with the same criterion. Sequence training was performed using the state-level Minimum Bayes' Risk (sMBR) criterion [24, 25]. Evaluation was done using Minimum Bayes' Risk (MBR) decoding [7].

To map the frame posteriors between PDTs, the transformation matrices, $P(s|s^m)$, were computed using the tandem GMM alignments, with a small discount of $\nu = 10^{-4}$. The discount cannot be too large, as there are many logical contexts. The interpolation weights were all set as equal, $\lambda_m = \alpha_m = \frac{1}{M}$.

5.1. Ensemble performance

The first experiment assesses the gains that can be obtained from both the RI and RF ensemble methods. Each ensemble consisted of four sMBR-trained models with the same architecture. Combination was done at the hypothesis level using MBR combination decoding [7], and at the frame level through a linear average of observation pseudo likelihoods (2).

Table 1: *Ensemble WER (%) performance*

Ensemble method	Single model				Combined	
	mean	best	worst	std dev	hypothesis	frame
Tok Pisin VLLP						
RI	47.8	47.6	48.0	0.18	46.3	46.7
RF	48.3	48.0	48.4	0.17	45.8	46.0
Javanese FLP						
RI	53.8	53.7	53.9	0.10	52.2	52.5
RF	54.1	54.0	54.3	0.14	52.3	52.4
HUB4						
RI	9.2	9.1	9.3	0.10	8.8	8.8
RF	9.3	9.2	9.4	0.10	8.7	8.7

The results in Table 1 show that significant combination gains can be achieved over single models by both ensemble methods. In the low resource tasks, the RF ensembles have worse performing single models, as their PDTs are less optimal. Despite this in Tok Pisin, the combined RF ensemble is able to significantly outperform the combined RI ensemble. In both Javanese and HUB4, the combined RF ensembles are able to match the RI ensemble performances, but not significantly outperform them. Perhaps the diversity between PDTs becomes less significant with larger PDTs. Using methods to encourage more PDT diversity [16, 17] may help. These results suggest that the RF method may be particularly helpful when the training data is extremely limited and the PDTs are small. It is also interesting to note that this trend is in spite of the Javanese and HUB4 RF ensembles having 102577 and 85840 RF tied states respectively, which is many more than for Tok Pisin, with 15094 RF tied states. The RF and RI methods provide different modes of diversity, and can be used concurrently to obtain a richer ensemble. It is therefore useful to be able to train a student toward teachers with different sets of state clusters.

Hypothesis combination outperforms frame combination in some of the ensembles, possibly because unlike frame combination, it does not require all models to produce time-synchronous states. However, frame combination is less computationally expensive. Furthermore, frame combination is indicative of the quality of the target posteriors that are used to train the students, and it shall therefore be used in the further experiments.

Table 2: *Mean single model WER (%) with standard training*

Dataset	CE	+ sMBR
Tok Pisin VLLP	50.2	47.8
Javanese FLP	55.9	53.8
HUB4	10.0	9.2

5.2. Student-teacher training

The next experiment uses the proposed method to train students toward both types of ensembles. The students of both ensembles used the same greedy PDTs as the RI ensembles. As a baseline for comparison, Table 2 shows the mean performance of single models using these PDTs, with standard training, of which the sMBR results are a repetition from Table 1.

Table 3: *Student-teacher training*

Ensemble method	Student WER (%) ST	+ sMBR	Ensemble WER (%)
Tok Pisin VLLP			
RI	46.9	46.6	46.7
RF	47.3	46.6	46.0
Javanese FLP			
RI	52.4	51.6	52.5
RF	52.7	51.9	52.4
HUB4			
RI	8.9	8.8	8.8
RF	9.2	9.0	8.7

The student performances with ST training are shown in Table 3. Here, frame combination of the teacher ensembles provides an indication of how well the students can be expected to perform. The results show that the proposed method is able to train students toward RF ensembles, achieving better performances than standard CE training with hard targets in Table 2. Further sMBR training of the students brings additional gains, though not significantly for HUB4. However, there is a consistent performance loss between the RF ensembles and their students after only ST training. This leads to the RF students performing worse than the RI students in all datasets. This degradation may be caused by the posterior mapping (15) or the limited student complexity.

5.3. Student model output complexity

The proposed method gives the freedom to select the student's state clusters independently of the teachers'. The final experiment investigates training students with PDTs of various sizes toward the RF ensembles. Using a larger PDT increases the student's phonetic resolution, and potentially reduces any degradation of the target posteriors arising from the posterior mapping of (15). Table 4 shows the results for Tok Pisin, which Table 1 suggests operates in a data quantity and PDT size regime that is able to benefit most from the RF method. Here, the larger students used either a PDT with 1800 leaves trained using greedy splits, or the RF tied states. The intermediate PDT size of 1800 leaves was chosen, as this was about the largest that could be generated for this dataset without relaxing the likelihood improvement threshold. Using the RF tied states as the student DNN outputs is inspired by [26], and gives the student the same phonetic resolution as the RF ensemble.

The results show that increasing the number of leaves in the students' PDTs allows them to better capture the RF ensemble performance, thereby mitigating the degradation in the proposed method. The best student performance after only ST training is obtained when the student DNN directly uses the RF tied states as outputs, and this student is able to outperform the RI student in Table 3. This shows that the gain of the RF ensemble over the RI ensemble in this dataset can be propagated to the student. Further sMBR training of this RF tied state student gives the best single system performance, and is able to meet the ensemble performance. However, this requires a large number of parameters, which may present a hindrance when deploying the ASR system on devices with hardware limitations. A possible method of reducing the number of parameters is to force the output layer linear transformation to be low-rank [27]. These results suggest that students with more complex outputs may be required to effectively capture the RF ensemble behaviour, because of the nature of its diversity. In Javanese and HUB4, RF ensemble students with PDTs having 10000 leaves give WERs of 52.4 % and 9.0 % respectively after only ST training, showing consistent improvements with increased student output complexity.

Table 4: *RF ensemble students with larger PDTs, for Tok Pisin*

Student PDT size	Student WER (%) ST	+ sMBR	Ensemble WER (%)
1000	47.3	46.6	
1800	47.0	46.3	46.0
15094 (RF tied states)	46.6	46.0	

By allowing the state clusters of the student to be chosen independently of those of the teachers, the proposed method introduces the possibility of using students with greater output complexities. Although this increases the computational cost of decoding, a balance can be chosen to make it far less than decoding through the ensemble.

6. Conclusion

This paper presents a method to perform ST training when the student and teachers use different sets of state clusters. This is accomplished by minimising the KL-divergence between the logical context posteriors of the student and teachers. To compute the logical context posteriors, an approximation is made that the probability of a logical context is independent of the observation when given the state cluster. This allows the proposed method to be implemented by mapping the teachers' frame posteriors to the student's state clusters. The experiments show that the proposed method allows the student to learn from teachers with different PDTs. Although degradation is observed between the student and RF ensemble performances, the proposed method also allows the use of larger PDTs for the student, which has been shown to improve the student performance.

The proposed method expands the flexibility of the ST framework, allowing for different sets of state clusters between teachers, and also for the student's state clusters to be chosen independently of the teachers'. It is therefore possible to use richer ensembles with multiple forms of diversity, so long as a mapping can be computed from the teachers' frame posteriors to the student's state clusters.

7. References

[1] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, Philadelphia, USA, Aug 2006, pp. 535–541.

[2] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1910–1914.

[3] L. J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, Montréal, Canada, Dec 2014, pp. 2654–2662.

[4] J. H. M. Wong and M. J. F. Gales, "Sequence student-teacher training of deep neural networks," in *INTERSPEECH*, San Francisco, USA, Sep 2016, pp. 2761–2765.

[5] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, Aug 2000.

[6] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[7] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, Oct 2011.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun 2014.

[9] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug 1997.

[10] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1915–1919.

[11] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *ICASSP*, Philadelphia, USA, Mar 2005, pp. 197–200.

[12] J. Xue and Y. Zhao, "Random forests of phonetic decision trees for acoustic modeling in conversational speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 519–528, Mar 2008.

[13] O. Siohan and D. Rybach, "Multitask learning and system combination for automatic speech recognition," in *ASRU*, Scottsdale, USA, Dec 2015, pp. 589–595.

[14] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT*, Plainsboro, USA, Mar 1994, pp. 307–312.

[15] T. Zhao, Y. Zhao, and X. Chen, "Building an ensemble of CD-DNN-HMM acoustic model using random forests of phonetic decision trees," in *ISCSLP*, Singapore, Sep 2014, pp. 98–102.

[16] C. Breslin and M. J. F. Gales, "Complementary system generation using directed decision trees," in *ICASSP*, Honolulu, USA, Apr 2007, pp. 337–340.

[17] H. Xu, G. Chen, D. Povey, and S. Khudanpur, "Modeling phonetic context with non-random forests for speech recognition," in *INTERSPEECH*, Dresden, Germany, Sep 2015, pp. 2117–2121.

[18] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, no. 4, pp. 359–394, Oct 1999.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *ASRU*, Hawaii, USA, Dec 2011.

[20] M. P. Harper, "IARPA babel program," http://www.iarpa.gov/index.php/research-programs/babel, 2011.

[21] M. J. F. Gales, K. M. Knill, and A. Ragni, "Unicode-based graphemic systems for limited resource languages," in *ICASSP*, Brisbane, Australia, Apr 2015, pp. 5186–5190.

[22] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The development of the Cambridge University RT-04 diarisation system," in *Rich Transcription Workshop (RT-04f)*, Palisades, USA, Nov 2004.

[23] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *INTERSPEECH*, Dresden, Germany, Sep 2015, pp. 3660–3664.

[24] M. Gibson and T. Hain, "Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition," in *INTERSPEECH*, Pittsburgh, USA, Sep 2006, pp. 2406–2409.

[25] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to MPE for large scale discriminative training," in *ICASSP*, Honolulu, USA, Apr 2007, pp. 321–324.

[26] O. Siohan, "Sequence training of multi-task acoustic models using meta-state labels," in *ICASSP*, Shanghai, China, Mar 2016, pp. 5425–5429.

[27] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP*, Vancouver, Canada, May 2013, pp. 6655–6659.